Morphologie Constructionnelle et Traitement Automatique des Langues : le projet MorTAL*

Georgette Dal

Nabil Hathout

Fiammetta Namer

Les contributions au présent volume témoignent la plupart de connexions entre la morphologie dérivationnelle (ou constructionnelle si l'on adopte la terminologie de D. Corbin (cf. Corbin (à paraître)), et d'autres champs réputés internes à la linguistique théorique : sémantique lexicale, syntaxe ou phonologie.

L'article qui suit s'intéresse, lui, à la connexion entre la morphologie constructionnelle et le traitement automatique des langues (désormais, TAL), domaine dont la visée n'est pas théorique mais applicative.

Contrairement à Fradin (1994a) que nous utiliserons à l'occasion, notre contribution ne se veut pas seulement théorique. En effet, même si, dans la première section, nous présenterons rapidement chaque discipline en enchaînant par un bref état des lieux sur les relations qu'entretiennent la morphologie constructionnelle et le TAL, nous consacrerons la deuxième section de cet article à un projet de construction de bases de données constructionnelles pour le TAL. Après avoir présenté rapidement ce projet, nommé « MorTAL », nous exposerons quelques-uns des bénéfices que le TAL peut en tirer en nous focalisant sur deux de ses domaines d'application : la recherche d'information et la fouille de textes (les bénéfices que la linguistique théorique peut en tirer apparaîtront, eux, en filigrane).

* Nous remercions nos deux relecteurs anonymes, et avons tenté de profiter au mieux de leurs remarques.

1. MORPHOLOGIE CONSTRUCTIONNELLE ET TAL : PRESENTATION DES DOMAINES ET DE LEURS CONNEXIONS

1.1. *Morphologie constructionnelle théorique*

Il est inutile de s'attarder à présenter la morphologie constructionnelle théorique dans un numéro consacré aux connexions qu'entretient cette discipline avec d'autres domaines de recherche. Aussi nous en tiendrons-nous à une définition minimale en disant qu'il s'agit là du champ de la linguistique qui recherche et étudie les régularités et principes gouvernant la construction des unités lexicales dotées d'une structure et d'un sens construits.

Même si la mise au jour de ces régularités et principes peut être réinvestie dans des applications à visée pratique, ainsi définie, la morphologie constructionnelle se situe donc sur le terrain de la linguistique théorique, en ceci qu'elle se donne comme seul objectif la description d'un pan de la langue, celui des mots construits, sans préoccupations applicatives dans des domaines qui lui sont extérieurs.

1.2. *Traitement automatique des langues*

La présentation que nous ferons du TAL sera un peu plus circonstanciée. Elle se fonde assez largement sur Fuchs éd. (1993) et sur Bouillon (1998), auxquels nous renvoyons pour plus de détails.

1.2.1. *Objectif*

L'objectif (et l'objet) du TAL est la conception de programmes informatiques capables de traiter automatiquement de données linguistiques exprimées dans une langue dite « naturelle » – où *langue naturelle* s'oppose à *langage artificiel* (informatique, mathématique, logique, etc.) –, en vue d'une application donnée.

Les programmes informatiques que conçoit le TAL et leurs applications dans des tâches précises ne constituent toutefois que la partie émergée du domaine. En effet, l'élaboration de ces programmes suppose un travail en amont qui peut se situer sur plusieurs territoires :

– celui de la *linguistique théorique*. On recherche alors des modèles linguistiques aptes à décrire les phénomènes de langue en recourant à des outils formels, majoritairement empruntés à la logique et aux mathématiques (on citera ici des modèles syntaxiques comme HPSG ou LFG créés il y a maintenant vingt ans en alternative au modèle chomskien). Si ces modèles se prêtent bien à l'écriture de grammaires pour ordinateurs, ils revendiquent toutefois résolument le statut de modèles linguistiques à part entière, dans la mesure où ils requièrent au moins autant de connaissances linguistiques que de connaissances informatiques (pour une présentation détaillée de ces grammaires dites « d'unification », cf. Abeillé (1993)).

– celui de la *linguistique informatique*. On se consacre, dans ce domaine, à l'implémentation de modèles linguistiques. Ces modèles sont souvent conçus dans ce but (cf. ci-dessus) mais peuvent également avoir à l'origine une vocation purement théorique : c'est le cas, en particulier, du modèle du Gouvernement et du Liage (*Government and Binding theory*), qui a servi de base à des applications informatiques (analyseurs syntaxiques, systèmes de génération de textes). La réalisation de ces applications requiert :
(i) soit l'utilisation d'une plate-forme de développement spécialisée (comme KIMMO [1] en morphologie, G-TAG [2] en génération de textes, ALEP [3] en analyse syntaxique), intégrant souvent des facilités, comme par exemple la représentation graphique des résultats, et masquant à l'utilisateur les aspects spécifiques aux langages de programmation,

---

[1] Le système KIMMO, qui se réclame de la morphologie à deux niveaux, est dû à Koskenniemi (1983). Ses avantages et limites ont été exposés dans Sproat (1992) et Fradin (1994a).

[2] G-TAG (cf. Danlos (2000)) est un système de génération automatique de textes, réalisé sur le modèle des *Grammaires d'Arbres Adjoints* (TAG), conçu par A. Joshi (1985). L'entrée du générateur G-TAG est un graphe événementiel, constitué interactivement au moyen d'un formulaire rempli par l'utilisateur, ce qui garantit la complétude des informations nécessaires à la synthèse du document.

[3] ALEP, « Advanced Linguistic Engineering Platform », est un environnement de développement pour l'analyse basée sur l'unification de structures de traits typées, conçu et développé à partir de 1996 d'après une initiative des communautés européennes.

(ii) soit celle d'un langage de programmation, ce qui constitue un environnement de travail de plus bas niveau. Certains langages soit plus appropriés que d'autres pour telle ou telle tâche du TAL : ainsi, on choisira Prolog pour l'analyse syntaxique (lorsqu'elle est compositionnelle et si les représentations sont arborescentes), et Perl, Lex ou Yacc pour l'analyse morphologique.
Comme l'indique sa dénomination *linguistique informatique*, ce domaine de recherche requiert des connaissances informatiques, mais aussi et d'abord des connaissances linguistiques : l'informatique est ici au service de la linguistique, comme support pour la description des phénomènes linguistiques, la vérification des hypothèses linguistiques, la validation des théories, etc.

– celui de l'*informatique linguistique*. Il s'agit également d'implémentation mais, cette fois, la part des connaissances linguistiques préalables à l'implémentation est réduite, voire inexistante. On se fonde en effet davantage sur les statistiques, les mathématiques ou sur la logique que sur la linguistique, et on fait l'hypothèse que les connaissances linguistiques peuvent être apprises automatiquement à partir de fragments en langue naturelle en entrée (à condition qu'ils soient minimalement annotés, le plus souvent).

Le TAL a donc à voir avec la linguistique théorique puisque, comme elle, il traite de données de langue. Mais, contrairement à elle, il ne manipule pas les données linguistiques pour elles-mêmes, et ne s'y intéresse que pour autant qu'elles entrent dans la conception de produits informatiques (commerciaux ou de recherche) dédiés à une tâche donnée. En outre, et contrairement à la vision parfois réductrice que peuvent en avoir certains linguistes théoriciens faute tout simplement de connaissance sur le domaine, mener des recherches dans le champ du TAL n'implique ni nécessairement être informaticien aguerri, ni nécessairement pratiquer une linguistique aspartamisée.

1.2.2. *La langue vue par le TAL*

De ce qui précède, il ressort que le TAL voit les langues naturelles comme des entités pouvant donner lieu à des calculs (y compris sémantiques : cf. Nazarenko éd. (1998) consacré au principe de compositionnalité et à sa mise en œuvre dans le

domaine du TAL), que ces calculs soient effectués en amont de l'implémentation ou en aval. Il rejoint en cela les préoccupations des linguistes théoriciens, dont les efforts tendent normalement à réduire au maximum les « exceptions », à débusquer les régularités derrière des irrégularités de façade, bref, à montrer que la langue est (au moins tendanciellement) ramenable à une grammaire au sens de "système" (sur les diverses acceptions du nom *grammaire*, cf. Flaux (1997)).

1.2.3. *Types de traitement*

On a coutume de distinguer deux grands types de traitement dans le domaine du TAL.

Traiter une entité, quelle qu'elle soit, suppose en effet deux états (un état initial E_1, un état final E_2) et un procès permettant la modification de E_1 en E_2. Ce qui est vrai en général vaut aussi pour le TAL, qui met en œuvre deux types de traitements susceptibles de se combiner. De façon très grossière :

– soit il prend en entrée une suite de lettres ou de sons, et produit en sortie un traitement linguistique de cette suite en vue d'une application donnée. Le traitement mené consiste alors en une analyse d'un fragment de langue naturelle (et en une production de données linguistiques). Comme l'objectif est applicatif et non théorique, on a coutume de nommer cette opération *analyse*,

– soit il prend en entrée des représentations formalisées de descriptions linguistiques (ces dernières pouvant elles-mêmes avoir été acquises automatiquement ou être issues de recherches théoriques), et produit en sortie un fragment de langue naturelle, aussi court soit-il (il peut s'agir, par exemple, d'une unité lexicale complexe). Le traitement mené consiste alors à générer un fragment de langue naturelle à partir de descriptions partielles de cette langue. Le TAL donne à cette opération le nom de *génération* (ou *synthèse*).

1.2.4. *Sous-domaines*

Les données linguistiques impliquées dans le TAL sont aussi diverses que celles sur lesquelles travaille la linguistique théorique : on peut ainsi opérer une partition dans le TAL selon les données manipulées, et distinguer la phonologie, la phonétique et la prosodie computationnelles, la morphologie

(flexionnelle et constructionnelle) computationnelle, la syntaxe computationnelle, la sémantique computationnelle (ce sont d'ailleurs ces diverses composantes qui structurent la première partie de Fuchs éd. (1993)), sans oublier l'analyse automatique du discours, à laquelle s'intéresse le TAL depuis longtemps déjà, même si l'intérêt apparaît sporadique [4].

1.2.5. *Domaines d'application*

Mais le TAL peut encore se définir par rapport au domaine d'application visé : sans prétendre à l'exhaustivité, on citera la recherche d'information [5], la traduction automatique, l'extraction d'information (selon le degré de sophistication, il peut s'agir d'une indexation automatique de document, ou d'un résumé automatique), la correction orthographique, le traitement de la parole, la production automatique de lettres (par exemple, réponse aux lettres de réclamation), de rapports, de commentaires (sport, météo, bourse, etc.), les systèmes de dialogue homme-machine (question-réponse finalisée) [6].

Chaque domaine d'application visé peut nécessiter (et nécessite souvent de fait) le recours à des données syntaxiques, des données morphologiques, des données sémantiques, etc.

1.3. *Morphologie constructionnelle et morphologie computationnelle*

1.3.1. *Des développements en parallèle*

De même que la morphologie constructionnelle a longtemps été le parent pauvre de la linguistique théorique, la prise en compte des données constructionnelles dans le domaine du TAL est assez récente : au niveau mondial, elle suit, avec un léger décalage peut-être, la courbe du développement des études théoriques en

---

[4] Cf. par exemple Pécheux (1969). Pour des travaux plus récents, cf. entre autres Clavier *& al*. (1995), Timimi (1999).

[5] Nous prenons ici l'option de décrire la recherche d'information comme un domaine d'application du TAL, même si, jusque très récemment, elle formait un domaine à part, avec une communauté de chercheurs bien distincte (comme le sont, du reste, en linguistique théorique, par exemple les syntacticiens ou les morphologues).

[6] Pour une présentation détaillée des domaines d'application du TAL et des enjeux (économiques, politiques et théoriques) associés aux industries de la langue, nous renvoyons à Fuchs éd. (1993 : 13-18).

morphologie constructionnelle (sur cette croissance en parallèle, cf. Fradin (1994b)).

Cependant, le français vient loin derrière les pays anglo-saxons, et ce aussi bien en morphologie théorique qu'en morphologie computationnelle. Plus encore peut-être que les autres langues, le français pâtit en effet de la réputation selon laquelle les unités lexicales construites y sont irrégulières sémantiquement, comme si le sens était une donnée observable tangible et non une construction théorique hypothétique nécessitant un travail d'abstraction non négligeable à partir de l'observable [7]. Aux exemples de cette réputation dans le champ de la morphologie théorique que l'on trouve un peu partout [8] font écho des assertions comme celle qui suit, empruntée, elle, à un ouvrage à orientation taliste (il s'agit de Bouillon (1998 : 48-49)). Elle reflète la vision que le TAL a des données constructionnelles, et notamment sa réticence à les traiter automatiquement [9]:

> Les informations dérivationnelles sont, par nature, moins régulières que les précédentes [i.e. les informations flexionnelles] et donc, moins bien adaptées à un traitement automatique :
> – tout d'abord, elles sont *peu systématiques* : une même séquence de caractères peut avoir différentes significations grammaticales et sémantiques. Ainsi, dans *anticonstitutionnel* ou *anticlérical*, *anti+* a le sens de *opposé à*, alors que *antimoine* ne signifie pas *opposé aux moines* (exemple tiré de Fuchs et al., 1993, p. 87). Et que dire de *chamelier* (qui garde des chameaux) et *chapelier* (qui vend des chapeaux) ? Ou encore *griserie* (construction de pierre de grès) et *plomberie* (qui a le sens d'industrie de la fabrication des objets en plomb) ?
> – ensuite, elles ne sont que *partiellement productives* : un nouveau mot ne va pas nécessairement subir les mêmes dérivations que les autres mots de sa classe. […] pour prendre des exemples français, pourquoi *décentrage* et *décentrement,* mais *centrage* et non **centrement* […] Ou *mangeable* et non **comportable* ?

---

[7] Cf. par exemple Aronoff & Anshen (1998 : 242) qui, après avoir souligné que « [c]omplex words often have conventional senses that differ slightly from their predicted sense », enchaînent avec les irrégularités phonologiques, « a little harder to detect ». On en infère que le sens, lui, n'est pas difficile à capter.

[8] Y compris dans des articles de vulgarisation, par exemple dans l'*EU* : « [...] dans certains cas [...] le contenu du mot se reflète en partie dans sa constitution en unités plus petites, mais douées de sens, que l'on appelle « morphèmes » : un mot comme « in-dé-racin-able » est ainsi constitué de quatre morphèmes, son sens global étant la résultante du sens de ses éléments constitutifs ; seuls les mots ainsi « motivés » morphologiquement pourraient dans cette perspective se prêter à une description sémantique –attitude qui [...] s'expose à bien des déboires, étant donné la fantaisie qui règne dans la morphologie du français » (Kerbrat-Orecchioni (1995)).

[9] Cette citation suscite de nombreux commentaires que nous ne ferons pas. Nous nous contenterons de renvoyer à Corbin & Corbin (1991) pour une réponse aux exemples en *–ier* donnés, et à Temple (1996) pour une réponse aux exemples en *–erie*.

– enfin, elles changent la catégorie du mot […]

1.3.2. *Le prototype de la morphologie computationnelle : la morphologie flexionnelle*

On s'aperçoit en outre que, quand elle est considérée comme possible, la prise en compte des données constructionnelles dans le domaine du TAL est plus souvent programmatique qu'effectivement réalisée.

Vu du côté du TAL, le champ de la morphologie a en effet un prototype : la flexion (certainement parce que cette dernière est, elle, considérée comme fondamentalement régulière [10]). Cette représentation prototypique a une conséquence sur la place laissée en TAL à la morphologie constructionnelle : soit on considère que la flexion sature le champ de la morphologie computationnelle (aucune place n'est alors laissée à la morphologie constructionnelle) ; soit on pose, par hypothèse, que la morphologie constructionnelle satisfait les mêmes principes que ceux qui prévalent en morphologie flexionnelle (la citation de P. Bouillon *supra* l'indique clairement), et, partant, que les données constructionnelles doivent se voir appliquer les mêmes algorithmes, les mêmes formalismes que la flexion. Plusieurs travaux, relatant l'implémentation de systèmes dédiés à la flexion, se terminent ainsi par l'annonce d'une transposition possible au traitement des unités lexicales construites (par exemple, Courtin *& al.* (1994)), annonce qui demeure le plus souvent programmatique.

1.3.3. *La prise en compte des données constructionnelles dans le domaine du TAL : état des lieux*

On vient de voir que, parce qu'il voit dans les données constructionnelles du français des phénomènes imprévisibles qui échappent en grande partie au calcul, le TAL s'y intéresse peu.

---

[10] … et pas seulement chez les talistes : cf. par exemple Stump (1998 : 17) : « Inflection is semantically more regular than derivation. […] Thus, the third-person singular present-tense suffix *–s* in *sings* has precisely the same semantic effect from one verb to the next, while the precise semantic effect of the verb-forming suffix *–ize* is somewhat variable (*winterize* 'prepare (something) for winter', *hospitalize* 'put (someone) into a hospital', *vaporize* '(cause to) become vapor') ».
Sur la régularité de la flexion opposée à la prétendue irrégularité de la dérivation, cf. Dal (2002).

Il arrive cependant que des systèmes automatiques intègrent une dimension constructionnelle, pour les raisons suivantes :
– la plupart du temps, l'analyse constructionnelle est envisagée en tant d'adjuvant à l'analyse syntaxique : *in fine*, l'analyse constructionnelle effectuée sert principalement, voire exclusivement, à l'étiquetage des mots inconnus (dont près d'un tiers est constitué d'unités lexicales construites, d'après plusieurs études), l'objectif étant de calculer automatiquement l'appartenance catégorielle de l'unité en cause, de façon à ne pas bloquer l'analyse syntaxique (cf. Fuchs éd. (1993 : 92), Laporte (1997 : 52)) ;
– l'objectif visé peut aussi être la structuration automatique d'une langue de spécialité. C'est par exemple le but explicite que s'assigne P. Zweigenbaum dans les travaux qu'il mène avec N. Grabar sur la langue médicale (cf. entre autres Grabar & Zweigenbaum (1999)) ;
– un troisième objectif récurrent assigné à la prise en compte par le TAL des données constructionnelles réside dans l'acquisition automatique de familles, avec des visées applicatives telles que la recherche d'information.

Il est possible que la liste qui précède ne soit pas exhaustive. Cependant, et en tout état de cause, le TAL minore les services que peut rendre l'intégration d'une dimension constructionnelle dans les systèmes qu'il conçoit. C'est ce que fera apparaître le point suivant.

2. LE PROJET MORTAL : PRESENTATION ET APPLICATIONS

Dans cette seconde partie, nous commencerons par présenter le projet MorTAL d'une façon délibérément non technique (pour une présentation technique, cf. Dal & al. (1999), et Hathout & al. (2002)). Nous verrons dans un second temps quels services peuvent en tirer la recherche d'information et la fouille de textes, qui sont deux domaines d'application du TAL en pleine expansion.

2.1. *Présentation*

Le projet MorTAL (acronyme de « MORphologie pour le TAL »), financé pour 3 ans par le Ministère français de la Recherche dans le cadre des actions concertées incitatives, réunit les co-auteurs du présent article, ainsi que Ch. Jacquemin. MorTAL est officiellement terminé depuis le 31 décembre 2002, mais nous continuons à y travailler à ce jour.

MorTAL est résolument orienté vers le traitement automatique, puisqu'il a d'abord été conçu pour combler le manque qu'on vient de souligner : il se veut d'abord un outil utile au (et, partant, utilisable par le) TAL (il n'a par conséquent pas un degré de sophistication dans l'analyse comparable à celui du projet expérimental de *Dictionnaire constructionnel du français*, lancé à SILEX par D. et P. Corbin [11]).

Du point de vue de la morphologie constructionnelle théorique, MorTAL constitue cependant aussi une gageure dans la mesure où l'élaboration d'un programme informatique capable d'analyser de façon (semi-)automatique et de générer automatiquement les unités lexicales construites du français présuppose et, en retour, prouve leur régularité.

La partie émergée du projet a la forme d'une base de données constructionnelles enrichies d'informations sémantiques, compilant une partie des entrées lexicales majeures (noms, adjectifs, verbes, adverbes) figurant dans *TLFnome* [12] et dans le *Robert électronique* (1994). A ce jour, le système d'analyse effectue l'analyse morphologique et sémantique complète des unités lexicales suffixées par *-able*, *-et(te)*, *-eur*, *-ifi(er)*, *-is(er),* *-ité*, *-ment*, *-tion*, soit au total environ 10 000 unités lexicales construites. Les procédés suivants sont, eux, partiellement couverts : suffixe *–oir*, préfixes *a-*, *dé-*, *in-* et *re-* (pour un total 4 000 unités lexicales), ainsi qu'une partie des convertis et des composés.

L'implémentation de la description des procédés constructionnels actuellement réalisée se fonde sur le modèle théorique de morphologie constructionnelle développé dans l'UMR SILEX par Danielle Corbin. MorTAL se situe donc dans

[11] Pour le dernier état du projet, cf. Corbin (1997).

[12] *TLFnome* est un lexique de formes fléchies construit à l'INaLF à partir de la nomenclature du *Trésor de la Langue Française*. Il contient actuellement 63 000 lemmes, 390 000 formes et 500 000 entrées. Il est complété par un second lexique de 36 400 lemmes supplémentaires issus de l'index du *TLF*.

la perspective de la linguistique informatique telle qu'elle a été définie *supra* (§ 1.2.1.).

Elle est réalisée grâce au système DériF (« Dérivation en Français ») qui, *modulo* les contraintes liées à la programmation, implémente les analyses théoriques formulées. Chacune des entrées traitées se voit ainsi associer :
(i) sa forme citationnelle subsumant conventionnellement toutes ses variantes flexionnelles,
(ii) sa catégorie lexicale,
(iii) une analyse constructionnelle sous la forme d'un schéma crocheté et étiqueté, reprise en clair dans un quatrième champ sous la forme d'une famille constructionnelle,
(iv) une glose formulée en langue (semi-)naturelle reflétant, pour les entrées construites, le résultat sémantique de l'application de l'opération constructionnelle la plus périphérique [13]. Dans le cas des unités lexicales non construites (signaler qu'une unité lexicale n'est pas construite est aussi une information constructionnelle), la glose duplique l'unité lexicale analysée.

Par exemple, l'adjectif construit *inarticulable* reçoit automatiquement la description suivante :

inarticulable/ADJ : [in [[articul(er) VBE] able ADJ] ADJ], (inarticulable/ADJ, articulable/ADJ, articuler/VBE) "qui n'est pas articulable"

tandis que le verbe non construit *articul(er)* est décrit par :

articuler/VBE : [articul(er) VERBE], (articuler/VBE) "articuler"

Cette première méthode, qu'on dira guidée linguistiquement, produit des analyses d'une bonne qualité linguistique (ses limites sont naturellement celles de la description et des choix théoriques faits en amont, combinées à celles liées aux contraintes de l'implémentation). Sa fiabilité linguistique a toutefois comme contrepartie un coût en temps important : chaque implémentation requiert en effet préalablement une analyse théorique du procédé concerné [14], une traduction de cette analyse dans un langage de

[13] La glose peut paraître pauvre aux yeux d'un morphologue théoricien. Mais MorTAL est conçu pour être un outil exploitable en TAL : les gloses doivent donc être directement exploitables, et aussi proches que faire se peut d'une définition « naturelle ».

[14] Programmer par exemple l'analyseur pour qu'il analyse les verbes en *dé-* pour lesquels peuvent revendiquer le rôle de base à la fois un nom et un verbe

programmation suivie d'une application de ce programme au corpus d'unités commençant ou terminées par le procédé analysé (ou une suite de lettres homomorphe de ce procédé), une vérification manuelle de cette analyse automatique suivie d'un amendement de l'implémentation, cette dernière phase pouvant se réitérer (pour les (rares) unités rétives à l'analyse menée, l'option a été faite de les traiter par listes d'exceptions : ainsi, sur 152 verbes en *–if(er)*, 6 sont considérés comme des exceptions, comme *qualifier* que l'on a décidé d'analyser comme non construit en français).

Aussi, pour pallier cet inconvénient et pour accroître la robustesse de la base, avons-nous exploré en parallèle une seconde piste pour construire notre lexique.

Ce second système, nommé DéCor (« Dérivation pour les Corpus »), se situe, lui, dans le paradigme de l'informatique linguistique (cf. *supra*, § 1.2.1.). Nous ne le développerons pas ici, pour deux raisons :

– d'une part, il n'associe pas de description sémantique aux données constructionnelles qu'il construit (il se limite à apparier les unités lexicales construites avec leurs bases présumées) : aussi l'exploitation des données en sortie est-elle moins riche pour le TAL que ne l'est celle des données fournies par DériF. Or, dans la suite de cet article, c'est de l'exploitation des données sémantiques qu'il s'agira surtout ;

– d'autre part, DéCor n'a en définitive pas à proprement parler fourni de données, sauf à titre de comparaison avec celles produites par DériF [15].

Nous dirons seulement :

---

demande des choix préalables et une véritable réflexion sur l'instruction sémantique du préfixe *dé-* : on en donnera un aperçu *infra*.

[15] Les résultats de DéCor ont été comparés avec ceux de DériF que nous avons considérés comme un étalon. Pour chaque suffixe, deux valeurs ont été calculées : le rappel (i.e. le rapport entre le nombre de résultats corrects de DéCor et le nombre total de résultats produits par DériF), la précision (i.e. le rapport entre le nombre de résultats corrects de DéCor et le nombre total de ses résultats). Il ressort des comparaisons effectuées que la qualité linguistique des appariements produits par DéCor dépend fortement de l'homogénéité de l'ensemble des bases potentielles. Ainsi, pour le suffixe *-able*, les résultats sont relativement bons (rappel et précision sont supérieurs à 95%) car les bases sont en très grande majorité des verbes, formellement différenciables des noms. Pour le suffixe *–ité*, ils le sont un peu moins (rappel et précision sont de l'ordre de 90%), car, même si les bases sont majoritairement des adjectifs, les formes de ces derniers sont assez proches des formes nominales. Enfin, les résultats sont insuffisants pour le suffixe *–is(er)* (rappel et précision sont de l'ordre de 50%) du fait de la répartition plus équilibrée entre les bases adjectivales et nominales que l'on peut difficilement départager sans utiliser d'informations sémantiques.

– que DéCor se fonde sur l'hypothèse qu'une partie significative de la structure du lexique peut être récupérée à partir de la liste des unités lexicales qui le composent,
– qu'il est beaucoup plus rapide à mettre en œuvre que DériF parce que les appariements base (présumée) / unité lexicale (présumément) construite qu'il produit sont appris automatiquement et qu'ils ne requièrent aucune connaissance linguistique préalable (quand plusieurs candidats se présentent au rôle de base d'une unité lexicale donnée, entrent en jeu des mesures statistiques, dont la fréquence de type du schéma qui connecte l'unité lexicale à chacun des candidats),
– que les résultats produits par les deux systèmes constituent deux sous-bases indépendantes.

2.2. *Applications en recherche d'information et en fouille de textes*

Plusieurs domaines d'application sont susceptibles de tirer bénéfice d'une base de données constructionnelles comme MorTAL. Nous ne présenterons toutefois que deux des destinataires privilégiés de notre projet : la recherche d'information (désormais RI) et la fouille de textes (désormais FT). Cette décision est triplement motivée :
– tout d'abord, prétendre vouloir passer en revue toutes les applications relevant du TAL susceptibles d'exploiter les résultats de la morphologie constructionnelle conduirait à appauvrir notre démonstration, par le passage d'un véritable raisonnement à ce qui ressemblerait plutôt à une liste à la Prévert,
– ensuite, nous avons orienté notre choix vers les domaines pour lesquels les unités lexicales et pluri-lexicales constituent un objet central,
– enfin, dans ce sous-ensemble d'applications, nous avons arrêté notre choix sur les domaines qui nous ont semblé répondre à des besoins émergents. Ainsi, la RI et la FT ont vu leur importance se développer avec l'augmentation exponentielle des besoins en recherche documentaire, liée à l'essor massif, ces dernières années, de la « société de l'information » : sous ce terme se cache la nécessité, tant pour les entreprises que pour les particuliers, de développer des outils leur permettant de maîtriser le contenu des documents, quelle qu'en soit la langue de rédaction, et quel que soit le format dans lequel il est stocké. Ces technologies, dont le développement a été accéléré par la généralisation des échanges

électroniques, sont liées d'une part à la recherche des informations contenues dans des documents (la RI), et d'autre part à la découverte de connaissances nouvelles à partir de ces informations (la FT).

Les objectifs de la FT et de la RI sont donc très différents : la FT implique que l'on ne connaît pas à l'avance les informations qui seront extraites des données textuelles ; rechercher les documents qui contiennent une information donnée suppose en revanche que l'utilisateur la connaisse. Il ne serait pas capable, sinon, de la formuler dans une requête (cf. Hearst (1999)). FT et RI constituent cependant des domaines connexes puisque les résultats de certains des traitements effectués en FT permettent d'améliorer le processus d'accès à l'information. Ainsi, la classification des documents en fonction des thématiques qui y sont développées conduit à une présentation plus synthétique des résultats ; la génération automatique d'associations entre termes produit des liens qui peuvent être utilisés pour l'extension de requêtes ; l'analyse des citations cooccurrentes permet de déterminer les principaux thèmes abordés dans une collection de textes et ainsi d'identifier les documents les plus importants, etc.

Après avoir présenté brièvement ces deux applications (section 2.2.1), nous explorons à la section 2.2.2. les avantages qu'elles peuvent tirer des résultats de MorTAL.

2.2.1. *Présentation de la RI et de la FT*

2.2.1.1. *Recherche d'information*

La RI est un ensemble de techniques et d'outils informatiques dont la finalité initiale était bibliographique : il s'agissait d'aider les usagers à trouver, dans des fonds documentaires, les références concernant un thème particulier. L'amélioration des capacités de stockage des ordinateurs a changé la nature du problème, qui n'est désormais plus d'exploiter des notices bibliographiques mais de conserver et d'accéder directement aux informations textuelles contenues dans les documents qui constituent les fonds.

Le champ de la recherche d'information moderne couvre ainsi plus largement la catégorisation des documents textuels, leur classification, leur catalogage et l'accès à leur contenu.

Longtemps réservée à une petite communauté de spécialistes, la RI est aujourd'hui connue et utilisée par un public plus large à travers les moteurs de recherche sur Internet.

Le problème général de la RI est de retrouver dans un ensemble de documents ceux qui contiennent des informations qui constituent des réponses à la requête d'un utilisateur. Les systèmes de RI doivent ainsi disposer d'une représentation des informations contenues dans les fonds documentaires et d'une procédure permettant de déterminer leur pertinence comme réponses à une requête particulière. Idéalement, ces systèmes devraient « comprendre » les informations textuelles contenues dans les documents et les requêtes (ces dernières, généralement formulées en langue naturelle, peuvent être considérées comme des documents supplémentaires). Une telle compréhension est malheureusement hors de la portée des systèmes de TAL ; étant donné les volumes des documents, une analyse sémantique de leur contenu par des opérateurs humains n'est pas non plus envisageable. Les systèmes de RI peuvent donc au mieux calculer une approximation du sens de ces informations, et évaluer leur proximité avec celui de la requête, de façon à classer les documents en fonction de leur pertinence comme réponses à la requête.

Les réponses, souvent médiocres, des systèmes de RI comme les moteurs de recherche montrent que des progrès importants restent à faire.

L'utilisation de représentations linguistiquement motivées est l'une des pistes qui pourrait conduire à l'amélioration des résultats des systèmes de RI (cf. Strzalkowski (1999)). Bien que de nombreuses expériences d'intégration d'outils de TAL dans des systèmes de RI n'aient pas été concluantes, la question de l'apport du TAL à la RI reste ouverte.

Par exemple, les outils de RI incluent souvent des traitements superficiels de morphologie flexionnelle destinés à diminuer le silence (*i.e.* le nombre de documents pertinents qui ne sont pas ramenés), mais il s'agit seulement de réduire les formes fléchies d'une même unité lexicale à une même chaîne de caractères en enlevant systématiquement aux formes du document une sous-chaîne finale susceptible de constituer une marque flexionnelle. Cette technique, appelée « racinisation », réduit par exemple les formes du verbe *fonctionner* en /fonctionn/. La suppression des sous-chaînes finales peut provoquer des erreurs, comme le fait qu'une requête comportant l'adjectif *frais* ramène les documents qui contiennent une forme du verbe *fraiser* mais pas ceux qui

incluent *fraîche* et *fraîches*, leur chaîne réduite étant *a priori* /fraîch/. Namer (2000) traite en détail des problèmes liés à l'utilisation des raciniseurs en matière de flexion**,** et montre l'importance d'intégrer aux tâches de RI une analyse flexionnelle valide linguistiquement, comme celle que fournit l'outil FLEMM qu'elle a développé.

La racinisation est également utilisée pour traiter la morphologie constructionnelle, en réduisant à leur radicaux les formes des lexèmes qui comportent certains affixes, mais les résultats fournis sont peu satisfaisants, en particulier pour le lexique général (pour des exemples, cf. Namer & Dal (2000)). Sur ce plan, l'apport de la base de données constructionnelles qu'est MorTAL peut être décisif, comme nous le verrons au § 2.2.2.

2.2.1.2. *Fouille de Textes*

Dans les entreprises et, à plus forte raison, dans les centres de documentation, la masse des documents manipulés est de plus en plus importante, leur contenu étant en outre susceptible de subir des mises à jours fréquentes, surtout s'il s'agit de documents de maintenance, de modes d'emploi, etc. Ces documents doivent pouvoir être classés, hiérarchisés et indexés en fonction de leur contenu, de façon à permettre à un expert du domaine d'accéder rapidement à l'information la plus pertinente.

Considérons par exemple le centre de recherche d'un laboratoire pharmaceutique confronté au problème de l'encéphalite spongiforme bovine. Le centre reçoit quotidiennement des quantités importantes d'articles, revues, livres, résumés et autres documents consacrés à ce sujet. Grâce à la hiérarchisation des informations contenues dans ces documents, l'utilisateur peut accéder sélectivement aux ouvrages consacrés à l'historique de la maladie, ou à la localisation des derniers cas enregistrés, ou aux derniers progrès vétérinaires enregistrés, ou aux différentes techniques utilisées pour le dépistage, etc.

Un autre exemple est celui de la gestion du courrier électronique par le service clientèle d'une entreprise. Ce dernier reçoit trois grands types de messages : demandes d'information, exposition de problèmes d'utilisation du matériel acheté, réclamations. Classer ces messages en fonction de leur contenu permet de les traiter plus rapidement, et, de plus en plus souvent,

ils le sont au moyen d'une réponse personnalisée générée automatiquement.

La fouille de textes, connue aussi sous l'appellation *analyse de l'information dans les textes*, est une tâche nécessaire pour rendre possible la structuration et la classification de ces informations. La combinaison de plusieurs méthodes est utilisée afin d'éviter la dispersion de l'information en termes de variantes, et donc afin de lui trouver une forme canonique. Les informations pertinentes des documents ainsi factorisées – et donc les documents eux-mêmes – sont alors organisées en clusters, sortes de réseaux dans lesquels les liens, par conséquent aussi les objets regroupés, sont pondérés en fonction de coefficients qui mesurent leur proximité sémantique [16].

Pour conceptualiser l'information contenue dans les documents, les approches les plus utilisées à l'heure actuelle relèvent de techniques comme l'indexation, l'appariement de variantes syntaxiques de termes, la recherche de co-occurrences, etc. En d'autres termes, les applications en FT n'ont jusqu'ici pas mobilisé directement de connaissances reliées à la morphologie constructionnelle, si on fait abstraction des travaux sur les variantes terminologiques. Nous verrons *infra* comment cette tendance commence à s'inverser.

2.2.2. *Apport de MorTAL pour la RI et la FT*

La RI et la FT peuvent utiliser, selon leurs besoins spécifiques, tout ou partie des informations constructionnelles et sémantiques associées aux entrées de notre lexique (dont on rappelle qu'il constitue une sous-partie du lexique attesté dans les dictionnaires de langue générale actuels) : nous le montrerons en détail aux § 2.2.2.2. et 2.2.2.3.

Mais chacun de ces domaines peut aussi tirer bénéfice de la partie immergée de MorTAL : en effet, les règles linguistiques implémentées par le système DériF sont indifférentes au fait que l'unité lexicale à analyser figure ou non dans les dictionnaires. MorTAL fournit par conséquent un outil d'analyse (et de génération) d'unités lexicales hors dictionnaires, permettant par la même occasion de traiter à égalité les unités lexicales utilisées

[16] Cf. Toussaint *& al.* (2000) pour une présentation détaillée de la fouille de textes ; pour plus de détails sur la classification des textes en clusters, on pourra se reporter à Toussaint *& al*. (1998).

dans les textes, indépendamment de leur caractère attesté. C'est ce que nous allons commencer par montrer.

2.2.2.1. *Traitement à égalité des unités lexicales dans et hors dictionnaires*

En matière d'analyse lexicale, l'un des gros problèmes que rencontre le TAL est celui des unités lexicales absentes des dictionnaires (de langue ou liés au système) [17].

Or, si l'on se fonde sur Froissart & Lallich-Boidin (1996), dans près d'un tiers des cas, en français, ces mots hors dictionnaires sont des unités lexicales dotées d'une structure construite [18]. En voici quelques-uns que n'attestent ni le *RE*, ni le *TLF*, ni le *Nouveau Petit Robert* (désormais, *NPR*) qui, à eux trois, assurent pourtant une bonne couverture du lexique attesté du français synchronique (le premier est emprunté à l'*Encyclopædia Universalis* (= *EU*), le second aux archives en ligne du *Monde*, le troisième au *Nouvel Observateur* ; c'est nous qui soulignons) :

La théorie de l'information postule que, pour être détectable au mieux, un signal doit être émis dans une bande de fréquence extrêmement étroite. (*EU*, *s.v.* **exobiologie**)

Jordi Pujol restera aussi comme un homme-clé de la démocratie espagnole. Directeur adjoint de l'édition catalane d'El Pais, Xavier Vidal-Folch salue "son intervention durant le putsch du 23 février 1981 et sa contribution à la gouvernabilité de l'Espagne (*Le Monde*, article de R. Rivais, publié le 31 octobre 2003)

Ce qu'on peut encore manger sans risque. Le guide anti-vache folle (*Le Nouvel Observateur*, n° 1880, 16-22 nov 2000)

En plus d'être construites, ces unités lexicales relèvent majoritairement de langues dites de spécialité (technolectes scientifique, philosophique, médiatique, économique, etc.), si bien que donner des outils pour les répertorier et les analyser peut permettre d'enrichir une base terminologique. DériF constitue l'un de ces outils.

---

[17] L'autre gros problème est celui de l'ambiguïté, ou, pour le moins, de la polyréférence des unités lexicales (la source d'une polyréférence n'est pas toujours une ambiguïté, ainsi le nom *portable*, dont la polyréférence est imputable à un seul sens construit du mot).

[18] Si l'on exclut les erreurs typographiques et orthographiques, l'autre grand contingent des formes non reconnues est constitué de noms propres (cf. Maurel *& al.* (1996)) et de sigles.

Une expérience menée en ce sens est relatée dans Dal & Namer (2000) et Namer & Dal (2000). Nous la résumons ici.

Dans cette expérience, après avoir proposé une description théorique sommaire des suffixation par *-able*, *-ité* et *-is(er)*, et des combinaisons que ces suffixes autorisent deux à deux (des six combinaisons *a priori* envisageables, seules sont possibles la suffixation par *-able* de bases suffixées par *-is(er)*, et la suffixation par *-ité* de bases suffixées par *–able*, si, naturellement, les conditions sémantiques pesant sur les bases sont satisfaites) [19], nous avons soumis ces résultats théoriques à l'épreuve des faits (en choisissant délibérément des corpus non dictionnairiques). La description théorique étant confirmée par les observations faites en corpus, nous avons ensuite sélectionné dans notre lexique d'entrée (constitué, on le rappelle, des nomenclatures du *TLF* et du *RE*) les dérivés en *-able* et en *-is(er),* et leur avons concaténé, moyennant d'éventuels ajustements formels (dont l'apparition d'un /i/ épenthétique lors de l'application de *-ité* à une base en *-able*) respectivement les suffixes *-ité* et *-able*. Le corpus ainsi obtenu a enfin été nettoyé des unités lexicales présentes dans le lexique d'entrée. Ont été produits de la sorte, par exemple, les dérivés *abaissabilité*, *acceptabilité* ; *académisable*, *africanisable*. *Acceptabilité* et *académisable* étant attestés dans le *RE* et/ou dans le *TLF*, seuls ont été conservés *abaissabilité* et *africanisable* (notre générateur, nommé GéDériF, a produit de la sorte un ensemble de 2 691 unités lexicales construites absentes du lexique d'entrée, et néanmoins linguistiquement plausibles d'après les hypothèses linguistiques effectuées et les vérifications faites en corpus).

Nous avons ensuite appliqué DériF à ce corpus d'unités lexicales inventées conformément aux régularités constructionnelles observées. Le résultat final de ces diverses opérations a donc la forme d'une base de données constructionnelles absentes des dictionnaires assorties d'une analyse constructionnelle et d'une glose. Par exemple, *abaissabilité* reçoit la description :

---

[19] Les autres combinaisons sont impossibles, (i) soit catégoriellement (et donc aussi sémantiquement) : suffixation par *–able* de bases suffixées par *–ité* ; suffixation par *–is(er)* de bases suffixées par *–ité*, suffixation par *–ité* de bases suffixées par *–is(er)*, (ii) soit seulement sémantiquement : nous avons proposé d'expliquer la rareté (dans et hors dictionnaires) de mots en *Xabilis(er)* en invoquant un clash sémantique entre le caractère nécessairement acquis des propriétés qu'expriment les adjectifs que sélectionne *–is(er)* et le trait foncièrement endogène des propriétés exprimées par les adjectifs en *–able*.

abaissabilité/NOM : [[[abaiss(er) VBE] able ADJ] ité NOM] (abaissabilité, abaissable, abaisser) ‘Propriété de ce qui est abaissable’

La dernière phase de l’expérience a consisté en une évaluation quantitative des résultats permettant d’établir dans quelles proportions ces 2 691 termes inventés se retrouvent effectivement dans des documents (papier, sur support électronique ou en ligne) représentant différents genres textuels : toutes créations confondues, nous avons constaté une utilisation effective d’entre 15 et 22 % de nos créations (par exemple, *colorabilité*, attesté à quatre reprises dans l’*EU*, *moussabilité*, attesté dans la *Banque des Mots*, *abordabilité*, attesté dans 52 pages web).

L’expérience qui vient d’être sommairement relatée indique que l’analyseur DériF sous-jacent au lexique MorTAL résout en grande partie le problème des unités lexicales hors dictionnaires, dès lors qu’elles sont structurellement complexes.

Il s’ensuit que, grâce au système DériF, la RI et la FT peuvent utiliser la famille constructionnelle et la glose non seulement des unités traitées dans MorTAL, mais aussi d’unités lexicales structurellement complexes absentes de notre base :

– si l’unité lexicale est dans notre lexique, l’une et l’autre sont d’emblée utilisables,

– si elle n’y figure pas, DériF commence par lui associer une analyse constructionnelle ainsi qu'une glose paraphrasant le sens de l'unité lexicale construite par rapport à celui de sa base. Par ailleurs, à chaque fois que cela est possible, DériF calcule une liste de traits sémantiques grâce à l'exploitation de contraintes régulières exercées par les procédés de construction de mots sur les unités lexicales appropriées (cf. Namer, sous presse).

2.2.2.2. *Utilisation de la famille constructionnelle*

C’est essentiellement la recherche d’information qui tirera profit des familles constructionnelles associées aux entrées de MorTAL (ou aux unités absentes de MorTAL traitées par DériF).

En effet, une question transversale à plusieurs sous-domaines du TAL et, plus généralement, de l'ingénierie linguistique concerne le repérage des variantes dans les documents, c'est-à-dire le repérage des séquences textuelles qui expriment les mêmes informations sous des formes différentes.

Le problème se pose de façon cruciale en RI, où, par exemple, une requête qui comporte les mots *décoder*, *rapidement* et

séquences devrait ramener les documents qui contiennent des formes morphologiquement apparentées comme *décodage*, *décodeur*, *code* ; *rapide*, *rapidité* ; *séquençage*, *séquenceur*.

Pour l'heure, les systèmes de RI prennent en compte ces variantes constructionnelles en les ajoutant tout simplement à la requête. Une telle extension n'est naturellement possible que si l'on dispose d'une ressource fiable qui fournit les familles constructionnelles des formes présentes dans les documents.

Pour les termes spécialisés d'un domaine, des techniques d'apprentissage à partir d'un thésaurus peuvent être mises en œuvre comme celle que proposent Grabar et Zweigenbaum (1999). En revanche, pour les formes qui appartiennent à la langue générale, ce type d'apprentissage basé sur la racinisation n'est pas envisageable du fait du nombre important de leurres qu'elle contient :

– mots non construits contenant un pseudo-affixe comme *peuplier* qui, bien que comportant la séquence *–ier*, ne doit pas être réduite à /peupl/, ou *francium* qui n'appartient pas à la famille de *français*,

– bases supplétives comme pour *fruit* et *fructifier*,

– démotivation comme pour *chauffeur* ou *lunette* qui, en synchronie, n'appartiennent plus aux familles constructionnelles de *chauffer ou de lune*,

– etc.

La solution la plus appropriée est d'utiliser une base de données constructionnelles (désormais, BDC) [20] comme MorTAL qui permet de retrouver les familles à partir des composantes connexes extraites des chaînes constructionnelles associées aux entrées et dont les descriptions ont été validées par un opérateur humain. Ainsi, dans le cas de l'exemple donné en début de paragraphe d'une requête comportant *décoder*, *rapidement* et *séquences*, la RI peut exploiter les chaînes constructionnelles parenthétisées fournies par MorTAL (cf. § 2.1.), comportant l'un des trois termes de la requête. Par exemple :

coder/VBE :	(coder/VBE, code/NOM)
décodabilité/NOM :	(décodabilité/NOM, décodable/ADJ, décoder/VBE, code/NOM)
décodage/NOM :	(décodage/NOM, décoder/VBE, code/NOM)
décodeur/NOM :	(décodeur/NOM, décoder/VBE, code/NOM)

[20] L'utilisation d'une BDC améliore globalement les performances des systèmes de RI, quelle que soit la taille des requêtes, comme l'ont montré pour l'anglais les études de Hull & Grefenstette (1996) et de Daille & Jacquemin (1998).

rapidement/ADV : rapidement/ADV, rapide/ADJ)
rapidité/NOM : (rapidité/NOM, rapide/ADJ)

séquençage/NOM : (séquençage/NOM, séquencer/VBE, séquence/NOM)
séquenceur/NOM : (séquenceur/NOM, séquencer/VBE, séquence/NOM)

Les listes partageant le ou les derniers éléments décrivent la famille constructionnelle de mots construits apparentés. Plus la taille des sous-listes partagées est importante, plus l'apparentement est pertinent. Ainsi l'échantillon de la BDC ci-dessus donne lieu aux familles suivantes, utilisables par la RI pour étendre la requête :

famille1 décodabilité; décodable; décodage; décodeur; décoder; coder; code
famille2 rapidement; rapidité; rapide
famille3 séquençage; séquenceur; séquencer; séquence

En recourant aux familles ainsi calculées pour étendre la requête, on réduit le silence. En contrepartie, le bruit (*i.e.* nombre de documents non pertinents ramenés) peut toutefois augmenter, du fait de la polyréférence des termes initiaux ou ajoutés. Par exemple, une requête qui comporte le seul mot *fraiseur* sera étendue à l'aide des formes des unités lexicales *fraiser* et *fraise* qui, vraisemblablement, ramèneront des documents relatifs à la boucherie, à l'agriculture et à la pâtisserie. Dans la pratique, ce bruit est limité parce que les requêtes comportent en général plusieurs termes, et que la cooccurrence de ces derniers avec des formes qui leur sont morphologiquement apparentés constitue une contrainte suffisante pour garantir la préservation de leur sens. Par exemple, *code* peut désigner le texte d'un programme informatique, mais il est alors peu probable que *ADN* ou *séquençage* apparaissent également dans le document qui le contient, à moins que ce dernier ne concerne d'un programme d'analyse des séquences ADN.

Une autre utilisation immédiate de notre BDC par la RI consiste à exploiter le fait que les relations constructionnelles sont orientées. On peut ainsi limiter l'expansion des requêtes aux seules relations mot-base → mot-construit ou fixer un nombre maximal d'étapes de construction entre les mots de la requête initiale et ceux de sa version étendue.

2.2.2.3. *Utilisation de la glose*

Alors que seule la RI tirera bénéfice de l'utilisation de la famille constructionnelle, à la fois la RI et la FT pourront exploiter la glose associée aux entrées de notre base (ou aux unités lexicales hors base analysées par DériF). En effet, nous allons voir dans cette section que la glose exprimant la relation sémantique entre une unité lexicale construite et sa base (cf. § 2.1. l'exemple d'*inarticulable*) est une donnée dont l'exploitation peut se révéler précieuse, non seulement pour affiner et enrichir les taux de précision et/ou de rappel en recherche d'information, mais également pour vérifier et pondérer les liens qui associent les concepts dans les clusters des systèmes de fouille de textes.

a) Exploitation de la glose : sous quelle forme ?

Rappelons-le, la glose reflète une relation sémantique entre deux items lexicaux, et est exprimée en langue (semi-)naturelle. Pour être exploitable, elle nécessite toutefois au préalable une traduction sous la forme de fonctions logiques qui doit être formulée dans un modèle formel compatible avec le langage d'interrogation de base de données utilisé pour classer et rechercher les documents.

Le tableau suivant donne un aperçu de correspondances susceptibles d'être établies entre des procédés affixaux, des gloses et des relations logiques (dans ce tableau, *ULC* abrège "unité lexicale construite", *af.* "affixe" ; la quatrième colonne traduit et généralise, sous forme de représentations logiques prédicat-arguments, la glose que DériF associe à l'ULC en entrée, et qui apparaît dans la deuxième colonne (on rappelle que la glose traduit l'application du procédé le plus périphérique) ; quand la relation met en jeu un tiers (en général, l'objet du verbe construit) celui-ci est désigné par *OBJ*. Enfin, *Nrect.* se lit "nom recteur" (de l'adjectif)).

ULC	**af .**	**Glose**	**rel. log.**
cristallisabilité$_N$	*-ité*	N= "propriété d'être *cristallisable*$_A$"	N = Propriété_A
incapable$_A$	*in-*	A = "(Nrect) non *capable*$_A$"	A = non_A(Nrect)
martyris-$_V$	*-is-*	V = "traiter (OBJ) comme un *martyr*$_N$ "	V = TraiterComme_ N(OBJ)
dépucel-$_V$	*dé-*	V = "priver (OBJ) de son caractère *puceau*$_A$ "	V = PriverDuCaractère_A (OBJ)

vérifiable$_A$	*-able*	A = "(Nrect.) que l'on peut *vérifier*$_V$"	A = QuOnPeut_V(Nrect)

En menant l'analyse constructionnelle d'une unité lexicale construite jusqu'à l'obtention de son primitif, on dispose ainsi, de proche en proche, de la combinaison des gloses calculées à chaque étape, donc, par transposition, de leur représentation formelle.

Par exemple, l'analyse d'un nom comme *indélocalisabilité*, dont la forme met en jeu les cinq procédés cités, se résume au moyen de la chaîne constructionnelle suivante :
(*indélocalisabilité, indélocalisable, délocalisable, délocaliser, local*)

Pour générer automatiquement cette chaîne, apparemment simple, DériF intègre les décisions linguistiques non triviales suivantes :

– ***indélocalisabilité / indélocalisable***
Indélocalisabilité met en jeu le préfixe *in-* et le suffixe *–ité*. On peut donc voir dans ce nom soit le produit de la préfixation par *in-* du nom *délocalisabilité*, soit le produit de la suffixation par *-ité* de l'adjectif *indélocalisable*.
Le lexique attesté compte quelques noms de propriété en *in-* nécessairement dérivés de noms (par exemple, *inharmonie*, *inconfort*). Ces dérivés expriment l'absence d'une propriété attendue, celle-là même que nomme leur base. L'existence de ces noms permet, en première analyse, de voir dans *indélocalisabilité* le résultat de la préfixation par *in-* du nom *délocalisabilité*. Ce n'est pas pourtant pas l'analyse qui a été retenue, pour des raisons sémantiques d'abord (le sens associé à cette construction, dont « absence de délocalisabilité » constitue une approximation, ne paraît pas pertinent), pour des raisons d'économie du système ensuite : si le préfixe *in-* forme quelques rares noms à partir de noms de propriété, il forme aussi de façon massive des adjectifs à partir d'adjectifs (*indirect*, *inexact*, etc.). Aussi avons-nous décidé que ce dernier cas constituerait l'analyse par défaut et, donc, que, quand un nom mettrait concomitamment en jeu le préfixe *in-* et un suffixe de nom de propriété, nous privilégierions pour *in-* l'analyse par défaut. DériF dérive donc *indélocalisabilité* de l'adjectif *indélocalisable* par suffixation par *-ité.*
– ***indélocalisable / délocalisable***

Indélocalisable, lui, est nécessairement construit sur l'adjectif *délocalisable*. Le préfixe *in-*, quand il met en exergue l'absence d'une propriété sélectionne en effet massivement sur des adjectifs (et, dans une moindre mesure on vient de le voir, des noms). *Indélocalisable* ne peut donc pas être dérivé d'un quelconque **indélocalis-*, qui constituerait en tout état de cause une aberration sémantique et catégorielle.

– ***délocalisable / délocaliser***

Comme plus haut, il existe *a priori* deux analyses pour *délocalisable* : soit on le dérive de *localisable* par préfixation en *dé-*, soit on le dérive de *localis(er)* par suffixation en *-able*.

Le préfixe *dé-* est susceptible de sélectionner des types de bases catégoriellement divers, cette variété catégorielle se retrouvant également dans les dérivés formés : verbes ou noms dénominaux (*bourse / débours(er)* ; *harmonie / désharmonie*), verbes déverbaux ou désadjectivaux (*fig(er) / défig(er)* ; *niais / déniai(ser)*), adjectifs désadjectivaux (*loyal / déloyal*). La première analyse est donc jusqu'ici licite, même si, statistiquement, le cas est rare. Elle pose en revanche un problème sémantique. En effet, dans la majorité des cas, le dérivé est un verbe exprimant le procès consistant à dissocier, selon des cas prévisibles, (i) du référent de la base (de r(B)), le référent du COD du dérivé (r(COD)) : *débarquer des marchandises* = /enlever r(*marchandises*) de r(*barque*))/, (ii) du référent du COD du dérivé (de r(COD)), le référent de la base (r(B)) : *dénoyauter des cerises* = /enlever r(*noyaux*) de r(*cerises*)/ [21]. Les cas de verbes construits sur base adjectivale ou verbale peuvent être considérés comme des cas d'espèce de (ii), une propriété ou un état étant conceptualisables comme des parties d'un objet, de même que les cas où le dérivé est un nom ou un adjectif. On peut en effet considérer que le préfixe *dé-* présent dans la poignée d'adjectifs désadjectivaux attestés (par exemple, *déloyal*, dont la construction remonte au 12^{e} siècle) et de noms dénominaux (*désharmonie*) est le même que celui qui opère dans *débarqu(er)*, par exemple : dans *déloyal* et *désharmonie*, *dé-* peut être décrit comme marquant une distanciation par rapport aux propriétés-repères qu'expriment *loyal* et *harmonie*. Dériver *délocalisable* de

[21] Il est intéressant de noter que cette double possibilité se retrouve dans les dérivés en *en-* qui peuvent exprimer le procès de mettre (i) le référent de leur COD dans le référent de leur base (*emprisonner quelqu'un*), (ii) le référent de leur base dans le référent de leur COD (*empailler un animal*).

Pour une analyse moins grossière de la préfixation par *dé-*, cf. les travaux de F. Gerhard, en particulier Gerhard (2000).

localisable impliquerait donc que cet adjectif exprime une distanciation par rapport à la propriété-repère qu'exprime *localisable*. Or, tel n'est pas son sens. Dans ce cas également, c'est donc d'abord une raison sémantique doublée d'un argument statistique (*dé-* forme plus souvent des verbes que des adjectifs) qui explique que nous n'avons pas dérivé *délocalisable* de *localisable.* L'analyse alternative, elle, ne pose en revanche aucun problème : comme les adjectifs en *–able* dérivés de verbes, *délocalisable* dit régulièrement des référents des arguments externes du verbe dont il dérive qu'ils présentent l'aptitude à être délocalisés.

– délocaliser / local

L'implémentation que nous avons faite de la représentation du préfixe *dé-* (et l'analyse linguistique préalable dont nous venons de donner un aperçu) nous ont conduits à interpréter, à chaque fois que cela est possible, un verbe en *déX* comme exprimant un procès annulant chez le référent de son objet la propriété exprimée par *X*. C'est ainsi que *délocaliser*, glosable par "faire perdre (au référent de l'objet du verbe) son caractère local", a selon nous pour base l'adjectif *local*, et non pas le verbe *localis-* : en d'autres termes, l'analyseur tient compte du fait que *-is-*, ici, fonctionne non pas comme suffixe (il n'est porteur d'aucune instruction sémantique) mais comme un marqueur de classe selon la terminologie utilisée notamment dans Corbin (1997), marquant iconiquement le fait que le résultat de la préfixation par *dé-* de *local* est un verbe de changement d'état.

Le nom *indélocalisabilité*, finalement, est associé à son primitif au moyen d'une combinaison de relations logiques. Cette combinaison résulte de la traduction compositionnelle de chacune des gloses appropriées dans le tableau ci-dessus (le symbole « _ », employé dans le parenthétisage, indique la position – non pertinente pour *indélocalisabilité* du nom recteur de l'adjectif *localisable*) :

indélocalisabilité =
propriété(non(QuOnPeut(PriverDuCaractère(local,OBJ), _)))

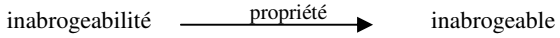
Outre l'obtention d'une représentation semi-formelle complète pour une unité lexicale construite [22], chaque relation prédicative

[22] L'obtention d'une représentation formelle complète pour la glose nécessite des informations (typage et aspect pour les procès, caractéristiques sémantiques pour l'ensemble des catégories lexicales) sur les entrées dont nous ne disposons pas sur les corpus d'entrée. Certaines de ces informations sont partiellement

peut être utilisée individuellement ou compositionnellement pour enrichir une hiérarchie de relations préexistantes.

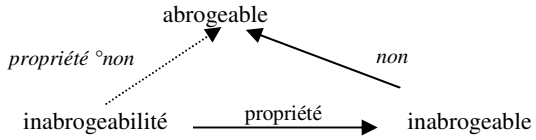
Supposons que le réseau sémantique utilisé dans une application de RI ou de FT code (au moins) les relations *contraire*, *hyponyme*, et *synonyme*. L'information fournie par la glose peut être vue comme l'expression de nouveaux prédicats reliant de nouveaux couples lexicaux.

C'est ce qu'illustre, pour l'entrée lexicale de *abroger*, l'insertion, dans la partie inférieure de la figure ci-dessous, des propriétés calculées à partir de notre base. Les termes sont reliés entre eux par des relations orientées, directes ou calculables. Chacune de ces relations orientées est représentée par une flèche étiquetée : les flèches en trait plein décrivent un mot *X* par sa relation directe avec *Y*. Ainsi, le graphe :



décrit la relation morphologique : *inabrogeabilité* = *propriété(inabrograble).*

Les flèches en trait pointillé décrivent, quant à elles, les compositions (notées '°') calculables par transitivité à partir des propriétés directes entre termes. Ainsi, dans :

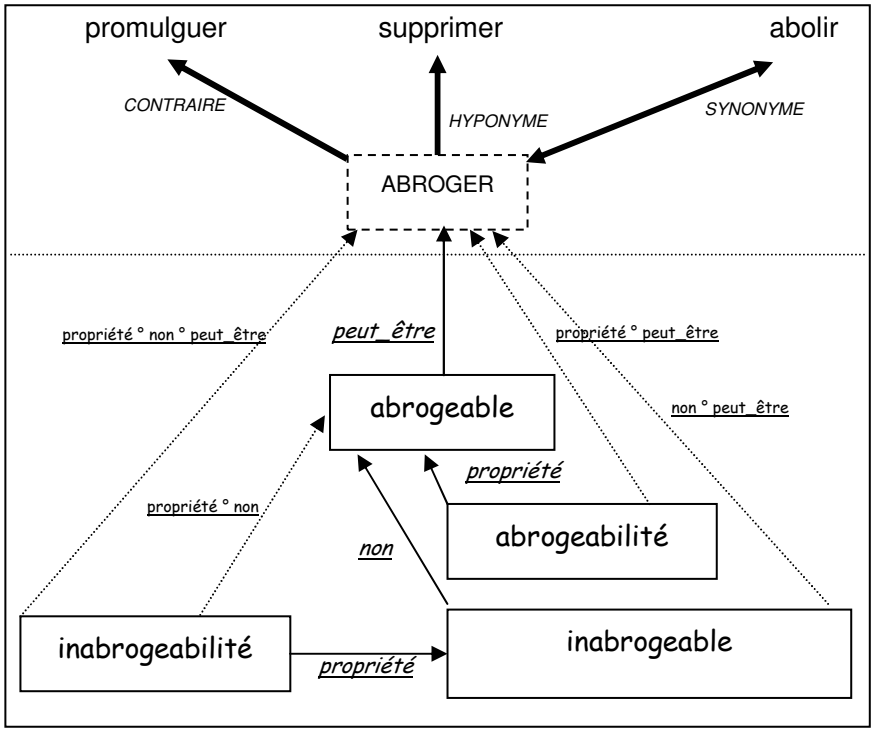


---

calculables (cf. notamment les travaux de M. Light – Light (1993) et (1996) –, pour ses expériences sur l'anglais). D'autres nécessitent un codage manuel. Pour l'ensemble de ces traits, néanmoins, la liste complète est impossible à dresser tant que l'ensemble des procédés constructionnels du français n'aura pas été étudié. A ce jour, plusieurs modèles formels ont été envisagés pour la représentation complète des gloses : citons, entre autres, la structure lexicale conceptuelle (CLS) de Jackendoff (1990) proposée pour la représentation des affixes constructionnels par notamment Lieber & Baayen (1993), ou encore le modèle en sémantique lexicale du « Lexique génératif » de Pustejovsky (1995), qui combine à la CLS d'un item lexical une représentation évènementielle et une représentation de la structure argumentale. Ce dernier modèle a récemment été expérimenté dans Namer & Jacquey (sous presse) pour la formalisation de la représentation sémantique des verbes dénominaux préfixés par *é-*, suivant l'analyse proposée dans Aurnague & Plénat (1997).

la relation *inabrogeabilité = propriété (non (abrogeable))* est la conséquence de la succession de liens directs entre *inabrogeabilité* et *inabrogeable* d'une part, et entre *inabrogeable* et *abrogeable* d'autre part.

Les différentes relations tissables autour de *abroger* fournissent ainsi le graphe suivant :



b) Exploitation de la glose : dans quel but ?

Nous venons de voir sous quelle forme exploiter la glose, voyons maintenant dans quel but.

Etant donné l'enrichissement d'un réseau par ces nouvelles relations issues de l'exploitation de notre base, on peut se demander comment celles-ci vont permettre effectivement

d'améliorer les résultats dans le cadre d'une application en fouille de textes et en recherche d'information.

Pour ce qui est de la FT, comme nous l'avons dit *supra*, cette discipline ne fait pas une utilisation directe, jusqu'ici, des résultats de la morphologie constructionnelle.

Une expérience, menée à Nancy (équipes Landisco/ATILF et Orpailleur/LORIA) dans le cadre du projet ILD&ISTC [23], cherche à déterminer le gain obtenu par l'intégration des informations sémantiques (reformulées) de notre base. Cette expérience vise à sélectionner les propriétés qui sont exploitables parmi celles que l'on peut formaliser à partir des gloses dans la base. Pour cela, on opère une indexation des textes à analyser au moyen des prédicats de la base, et on évalue leur impact dans la production des clusters, de façon à calculer le taux d'amélioration.

La RI peut, elle aussi, tirer partie des informations sémantiques qu'apportent les gloses de MorTAL pour indexer les documents de façon plus pertinente, mais également pour exploiter directement ces informations pour traiter des requêtes.

Considérons l'exemple suivant : « Comment fonctionne la xérographie ? ». Un document qui contiendrait le groupe *fonctionnement des copieurs xérographiques* constituerait une réponse appropriée à cette demande, tandis qu'un autre, qui traite de la « fonction du service de xérographie » ne conviendrait pas. Le fait que le premier document soit plus pertinent que le second peut être déterminé en utilisant la glose de l'entrée *fonctionnement* qui est du type « manière de fonctionner » [24], et qui répond directement à la question en *comment*. De même, la glose « accomplir une fonction » associée à *fonctionner* met en évidence le fait que le second document ne répond pas aux éléments de sens *comment* et *accomplir*. Disposer d'une description suffisamment fine du sens des mots construits permet ainsi de réhabiliter les opérateurs sémantiques, habituellement considérés en RI comme insuffisamment discriminants pour être utilisés comme index ou pour l'extension des requêtes.

---

[23] « Ingénierie des Langues, du Document et de l'Information Scientifique, Technique et Culturelle », projet soutenu par le MEN dans le cadre du Contrat de Plan Etat Région.

[24] La paraphrase « manière de fonctionner » associée de façon récurrente au nom *fonctionnement* est précisée dans la « liste d'exceptions » (la glose attendue pour les noms en *–ment* dérivés de verbes est par défaut « action, résultat de l'action de V »).

La RI peut également utiliser les gloses pour contrôler les variations morpho-syntaxiques. Ce problème se pose de manière cruciale lorsque l'on indexe les documents plus finement, en utilisant les termes du domaine plutôt que la liste de leurs formes. L'hypothèse sous-jacente est que l'utilisation d'index plus fins permet de mesurer plus précisément la proximité sémantique d'une requête avec les documents des fonds. Malheureusement, un même concept peut s'actualiser de plusieurs manières : par exemple, celui qui est associé au terme *mesure de la concentration atmosphérique en CO2* peut aussi se réaliser comme un groupe verbal /mesurer la concentration du CO2 dans l'atmosphère/, ou comme une proposition /le CO2 concentré dans l'atmosphère est mesuré/.

La principale difficulté à résoudre lors du repérage des variantes d'un terme est de déterminer l'acceptabilité de ces dernières. Il existe en effet un continuum entre les variantes valides qui permettent de paraphraser le terme initial (*baisse de la température de surface du support* : /la température de surface du support baisse/) et les variantes impropres (*développement d'application* : */appliqué au développement/) qui passe par des réalisations de concepts plus spécifiques (*méthode de conception* : ? *méthode de conception détaillée*), de concepts complémentaires (*mise en circulation* : * *mise hors circulation*), etc. Fabre (1998), à qui sont empruntés ces exemples, présente une étude détaillée des variations [terme nominal / groupe verbal] qui montre que la préservation des relations argumentales entre les constituants du terme initial est un critère opératoire qui permet de distinguer les variantes correctes de celles qui ne le sont pas. Il permet d'affiner les méta-règles de l'analyseur FASTR (Jacquemin (1997)), en prenant en compte la voix et la valence des verbes et le type constructionnel des noms déverbaux (agentif ou processif) [25]. Cette dernière information, codée manuellement dans cette expérience, pourra être calculée facilement à partir de la glose associée à ces noms dans la base MorTAL : leurs représentations formelles sont du type N=qui_exerce_l'activité(V) dans le premier cas (exemple : activateur = qui_exerce_l'activité(*activer*)), et N=action_de(V) dans le second (exemple : action_de(*augmenter*)). Cette information permet d'associer les traits [+ agentif] à *activateur* et [+ processif] à *augmentation*. Il devient alors possible de distinguer à l'aide de la méta-règle R deux variations comme :

[25] Une évaluation de ces méta-règles est présentée dans Fabre & Jacquemin (2000).

activateur de régénération cellulaire : */la régénération cellulaire active/ ; *augmentation de l'intensité lumineuse* : /l'intensité lumineuse augmente/. R ne reconnaît que la seconde variation.

R : N1 de Dét? N2 A3 : Dét N2 A3 V1
<N1 base> = V1 ; <N1 type> = processif

L'utilisation des gloses permet également de repérer des variantes qui sont pour l'instant éliminées car leur traitement génère trop de résultats incorrects. Par exemple, elles permettent de contrôler l'application d'une méta-règle destinée à repérer des constructions attributives comme : *l'indéformabilité de l'ensemble de la voilure* : *l'ensemble de la voilure est indéformable* en tenant compte du fait que *indéformabilité* exprime une propriété. En revanche, cette méta-règle ne s'appliquera pas à *l'économie des moyens de transport* : **ces moyens de transport sont économes* puisque *économie* n'a pas une glose du type propriété(*économe*), mais plus certainement du type propriété(*économique*).

EN GUISE DE CONCLUSION…

Il est inutile de souligner que l'exposé d'une recherche est souvent obsolète quand il vient à parution : c'est en général vrai en linguistique théorique, cela l'est encore davantage quand, comme ici, il s'agit d'une expérience applicative en cours de réalisation. Aussi MorTAL aura-t-il peut-être beaucoup évolué d'ici à ce que le présent travail paraisse, pas seulement quant à son contenu (l'évolution est, en la matière, prévisible), mais quant à sa conception (il a du reste connu plusieurs états depuis sa première version) ; peut-être aurons-nous d'ici là pensé à d'autres applications possibles ; peut-être même nous serons-nous aperçus que les applications que nous avons ici exposées sont en définitive peu pertinentes, non pas quant à leur intérêt (nous sommes convaincus qu'elles sont linguistiquement intéressantes), mais eu égard au gain obtenu pour les domaines d'application concernés. Seul l'avenir nous le dira, mais, en tout état de cause, l'expérience relatée ne peut qu'être enrichissante aussi bien pour le TAL que pour la théorie : pour le TAL parce qu'au moins, elle fait apparaître que les informations constructionnelles ne se limitent pas à des appariements formels, pour la théorie parce

qu'elle confronte des hypothèses théoriques aux données empiriques.

GEORGETTE DAL
U.M.R. « SILEX »
Université de Lille 3 & CNRS
NABIL HATHOUT
U.M.R. « ERSS »
Université de Toulouse-le Mirail & CNRS
FIAMMETTA NAMER
LANDISCO / U.M.R. « ATILF »
Université de Nancy 2 & CNRS

BIBLIOGRAPHIE

ABEILLÉ A. (1993), *Les nouvelles syntaxes. Grammaires d'unification et analyse du français*, Paris, Armand Colin.

ARONOFF M. & ANSHEN F. (1998), « Morphology and the Lexicon : Lexicalization and Productivity », *in* A. Spencer & A.M. Zwicky eds, *The Handbook of Morphology*, Blackwell Publishers, pp. 237-247.

AURNAGUE M. & PLENAT M. (1997), « Manifestations morphologiques de la relation d'attachement habituel », *in* D. Corbin, B. Fradin, B. Habert, F. Kerleroux & M. Plénat éds, *Silexicales* 1, pp. 15-24BOUILLON P. (1998), *Traitement automatique des langues naturelles*, Paris/Bruxelles, éditions Duculot.

CLAVIER V., LALLICH-BOIDIN G., ROUAULT J. & TIMIMI I. (1995), « Analyse Automatique du Discours, Perspectives 1995 », 3rd International Conference on Statistical Analysis of Textual Data JADT95, Rome.

CORBIN D. (1997), « La représentation d'une "famille" de mots dans le *Dictionnaire dérivationnel du français* et ses corrélats théoriques, méthodologiques et descriptifs », *Recherches Linguistiques de Vincennes* 26, Université de Paris VIII, pp. 5-37 + Supplément I-VIII.

(à paraître), *Le lexique construit. Méthodologie d'analyse*, Paris, Armand Colin.

CORBIN D. & CORBIN P. (1991), « Un traitement unifié du suffixe *-ier(e)* », *Lexique* 10, pp. 61-145.

COURTIN J., DUJARDIN D., GENTHIAL D. & KOWARSKI I. (1994), « Analyse et génération morphologique avec le système PILAF », *T.A.L.* 35/2, pp. 93-109.

DAILLE B. & JACQUEMIN C. (1998), « Lexical Database and Information Access: A Fruitful Association », *in Proceedings of the First*

International Conference on Language Resources and Evaluation, pp. 669–673, Granada, Espagne, ELRA.

DAL G. (2002), « A propos d'une idée reçue, ou de la prétendue irrégularité de la dérivation », *Bulag* 27, pp. 57-73.

DAL G., HATHOUT N. & NAMER F. (1999), « Construire un lexique dérivationnel : théorie et réalisations », *in Actes de la VI^e conférence sur le Traitement Automatique des Langues Naturelles (TALN'99), Institut d'Etudes Scientifiques de Cargèse, Corse, 12 - 17 juillet 1999*, pp. 115-124.

DAL G. & NAMER F. (2000), « Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations », *TAL* 41-2, pp. 423-446.

DANLOS L. (2000), « G-TAG : a Lexicalized Formalism for Text Generation inspired from Tree Adjoining Grammar : TAG issues », *in* A. Abeillé et O. Rambow eds., *Tree-adjoining Grammars*, Stanford, CSLI.

EU = CD-ROM *Encyclopædia Universalis*, version 2.0, Paris, Encyclopædia Universalis, 1995.

FABRE C. (1998), « Repérage de variantes dérivationnelles de termes », *Carnets de Grammaire* 3, ERSS (CNRS & Université de Toulouse-Le Mirail), Toulouse.

FABRE C. & JACQUEMIN C. (2000), « Boosting Variant Recognition with Light Semantics», in *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, pp. 264–270, Luxembourg.

FLAUX N. (1997), *La grammaire*, coll. « Que sais-je ? » 788, Paris, Presses Universitaires de France ; 1^e éd. : 1993.

FRADIN B. (1994a), « L'approche à deux niveaux en morphologie computationnelle et les développements récents en morphologie », *T.A.L.* 35/2, pp. 9-48.
(1994b), « Introduction », *T.A.L.* 35/2, pp. 3-7.

FROISSART C. & LALLICH-BOIDIN G. (1996), « Morphologie robuste et analyse automatique de la langue : étude réalisée à partir des corpus de l'évaluation GRACE », *in Actes du séminaire Lexique. Représentations et Outils pour les bases lexicales. Morphologie robuste*, Grenoble, CLIPS-IMAG, les 13 et 14 nov. 1996, pp. 88-96.

FUCHS C. éd. (1993), *Linguistique et traitements automatiques des langues*, Paris, Hachette supérieur.

GRABAR N. & ZWEIGENBAUM P. (1999), « Acquisition automatique de connaissances morphologiques sur le vocabulaire médical », *in Actes de la VI^e conférence sur le Traitement Automatique des Langues Naturelles (TALN'99), Institut d'Etudes Scientifiques de Cargèse, Corse, 12 -17 juillet 1999*, pp. 175-184.

HATHOUT N., NAMER F. & DAL G. (2002), « An experimental constructional database : the MorTAL project », *in* P. Boucher-ed, *Many morphologies*, Cambridge Mass., Cascadilla Press, pp. 178-209.

HEARST M. (1999), « Untangling Text Data Mining », *in Proceedings of the 37^th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, pp. 3-10.

HULL D. A.& GREFENSTETTE G. (1996), *A Detailed Analysis of English Stemming Algorithms*, Rapport technique MLTT-96 023, Rank Xerox Research Center, Meydan, France.

JACQUEMIN C. (1997), *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus,* Mémoire d'Habilitation à diriger des recherches, Université de Nantes.

JACKENDOFF R. (1990), *Semantic Structures*, Cambridge, MIT Press.

JOSHI A.K. (1985), « How much context-sensitivy is necessary for characterizing structural descriptions ?», *in* D. Dowty, L. Karttunen & A. Zwicky eds, *Natural Language Processing: Psycholinguistic, Computational and Theoritical Perspectives*, Cambridge, Cambridge University Press.

KERBRAT-ORECCHIONI C. (1995), article **Sémantique**, *in* CD-ROM *Encyclopædia Universalis*, version 2.0, Paris, Encyclopædia Universalis**.**

KOSKENNIEMI K. (1983), *Two-level Morphology*, University of Helsinki, Helsinki.

La banque des mots, revue de terminologie française publiée par le conseil international de la langue française, Paris, Conseil international de la langue française.

LAPORTE E. (1997), « Les mots. Un demi-siècle de traitement », *T.A.L.* 38/2, pp. 47-68.

LIEBER R. & BAAYEN R.H. (1993), « Verbal prefixes in Dutch : a study in lexical conceptual structure », *Yearbook of Morphology 1993*, pp. 51-78.

LIGHT M. (1993), « A Computational Theory of Lexical Relatedness », The University of Rochester Computer Science Department Rochester, New York 14627 Technical Report 421.
(1996), « Morphological Cues for Lexical Semantics »,' PhD. Thesis, Department of Computer Science, University of Rochester.

MAUREL D., BELLEIL Cl., EGGERT E. & PITON O. (1996), « Réalisation d'un dictionnaire électronique relationnel des noms propres du français », in *Actes du séminaire Lexique. Représentations et Outils pour les bases lexicales. Morphologie robuste*, Grenoble, CLIPS-IMAG, les 13 et 14 nov. 1996, pp. 164-175.

NAMER F. (2000), « FLEMM : un analyseur flexionnel du français à base de règles », *TAL* 41-2, pp. 523-549.
(sous presse), « Automatiser l'analyse morpho-sémantique non affixale : le système DériF », *Cahiers de Grammaire* 28.

NAMER F. & DAL G. (2000), « GéDériF: automatic generation and analysis of morphologically constructed lexical resources », *in Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May – 2 June 2000*, pp. 1447-1454.

NAMER F. & JACQUEY E. (sous presse), "Lexical semantics and derivational morphology: the case of the popular 'é-' prefixation in French"*, Proceedings of the 2nd International Workshop on Generative Approaches to the Lexicon*, Genève.

NAZARENKO A. éd. (1998), *T.A.L.* 39, 1, « Compositionnalité ».

NPR = Version électronique du Nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française, Nouvelle édition (version 2), Dictionnaires Le Robert / VUEF, 2001.

PECHEUX M. (1969), *Analyse automatique du discours*, Paris, Dunod.

PUSTEJOVSKY J. (1995), *The Generative Lexicon*, Cambridge, Mass., the MIT Press ; 2nd printing : 1996.

RE = Le Robert électronique, Disque optique compact CD-ROM, Paris, Dictionnaires Le Robert, 1994.

SPROAT R.W. (1992), *Morphology and Computation*, Cambridge, Massachusetts / London, England, The MIT Press.

STRZALKOWSKI T. (1999), « Preface », *in* T. Strzalkowski ed., *Natural Language Information Retrieval*, Kluwer Academic Publishers, Dordrecht., pp. xiii–xxii.

STUMP G.T. (1998), « Inflection », *in* Spencer A. & Zwicky A.M. eds, *The Handbook of Morphology*, Oxford / Malden, Mass., Blackwell Publishers, pp. 13-43.

TEMPLE M. (1996), *Pour une sémantique des mots construits*, Villeneuve d'Ascq, Presses Universitaires du Septentrion.

TIMIMI I. (1999), *De la paraphrase linguistique à la recherche d'information. Le système 3AD : théorie et implantation (aide à l'analyse automatique du discours)*, thèse pour l'obtention du diplôme de Doctorat d'informatique et communication, Université Stendhal, Grenoble.

TOUSSAINT Y., NAMER F., JACQUEMIN Ch., DAILLE B., ROYAUTE J. & HATHOUT N. (1998), « Une approche linguistique et statistique pour l'analyse de l'information en corpus », *Actes de la Conférence TALN'98*, Paris.

TOUSSAINT Y., SIMON A & CHERFI H (2000), « Apport de la fouille de données textuelles pour l'analyse de l'information », *in Actes de la Conférence IC'2000*, Toulouse.

TLF = Trésor de la langue française. Dictionnaire de la langue du 19e et du 20e siècle (1789-1960), 16 vol., Paris, Éditions du CNRS (t. 1-10) / Gallimard (depuis le t. 11), 1971-1994.